Clustering-based measurement of dependence

Raffaella Piccarreta, Marco Bonetti, Sergio Venturini
Istituto di Metodi Quantitativi
Università Bocconi
Viale Isonzo 25, 20137 Milano, Italy
{raffaella.piccarreta, marco.bonetti, sergio.venturini}@unibocconi.it

Abstract: A measure of the dependence of a multivariate response variable upon a categorical variable is introduced. Its characteristics are explored via simulations by referring to a specific mixture association model. Inferential aspects are investigated using a permutation test approach. We present preliminary results.

Keywords: Association, IGP, Permutation tests

1 Introduction

Kapp and Tibshirani (2007) introduce the IGP (In Group Proportion) measure within the context of validating clusters. Let P_T be a partition of N observations on a multivariate variable Y into K clusters. Here, T denotes the (training) data set. Suppose that new observations on Y are available in a second dataset D. It is of interest to assign them to one of the previously determined clusters; the IGP has been introduced to evaluate the adequacy of the chosen assignment procedure. (The classification procedure may be defined in different ways).
Let $C_T(i)$ indicate the cluster to which the i-th observation in D is assigned, with $i = 1,\ldots,N_D$ and N_D indicating the size of D. Denote by P_D the resulting partition of D.
The (overall) IGP is defined as the proportion of cases in D that are classified to the same group as their nearest neighbor. More precisely, let $NN(i) \in D$ indicate the nearest neighbor of the i–th observation in D, and let $C_T(NN(i))$ denote the cluster to which $NN(i)$ is assigned. The IGP is therefore $IGP(P_D) = \frac{1}{N_D}\sum_{i=1}^{N_D} 1[C_T(NN(i)) = C_T(i)]$.
We propose to use the IGP index to measure the extent of the association between one set of response variables, Y, and one or more explanatory variables X, both observed on a *single* dataset. In particular, we focus on the case of one categorical variable X, taking values $x_1^*,\ldots,x_K^*$. Let P_X indicate the partition induced by these K groups and $C_X(i)$ the group to which the i–th observation belongs, where $C_X(i) = k$ if $x_i = x_k^*$. We define:

$$IGP(Y|X) = \frac{1}{N}\sum_{i=1}^{N} 1[C_X(NN(i)) = C_X(i)] \tag{1}$$

If the responses Y are related to X, then P_X should provide a good partition also with respect to Y, characterized by a high value of $IGP(Y|X)$.
In their paper, Kapp and Tibshirani consider two different data sets. The first one, T, is used to find clusters, and the observations of the second one, D, are assigned to those clusters. The IGP measures the reliability of this procedure, i.e., if and to which extent the clusters obtained on T provide an adequate prediction of cases in D. The two data

sets contain information on the same variables. In other words, an overall data set is partitioned by row into one training and one validation set.
In our problem the *same* data set is used but it is partitioned by columns. This can be described as a sort of nonparametric ANOVA problem on possibly multivariate responses. Below we explore the main features of this IGP-based approach through simulations based on a specific multivariate association model, with particular attention to the inferential aspects of the approach.

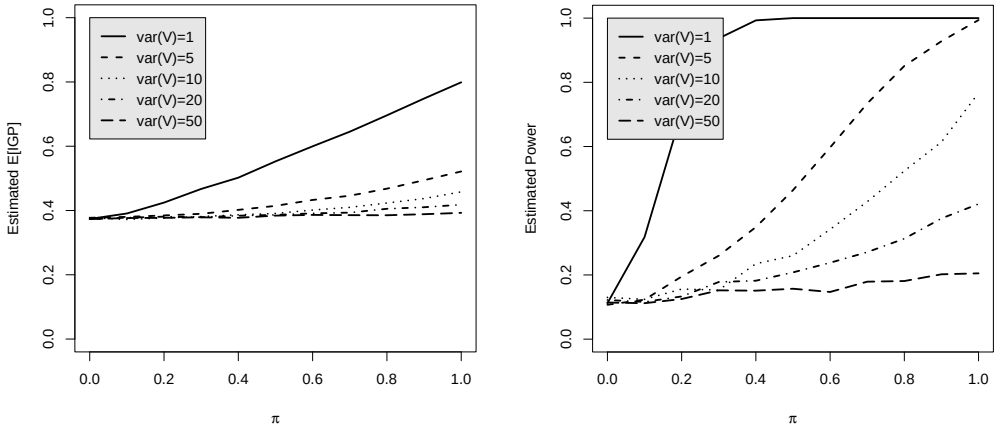
2 IGP as a measure of association

Consider the association model between X and Y such that Y is a mixture of two distributions f_V and f_Z with mixing parameter $\pi \in [0,1]$. Also, V and Z are distributed as two mixtures, each of three components (f_1^V, f_2^V, f_3^V) and (f_1^Z, f_2^Z, f_3^Z) respectively, with mixing vectors $(\alpha_1^V, \alpha_2^V, \alpha_3^V)$ and $(\alpha_1^Z, \alpha_2^Z, \alpha_3^Z)$. V and Z are assumed independent.
In particular, we set

$$\begin{array}{lll} f_1^V \sim N((0,2)^T, \sigma_V^2 I) & f_2^V \sim N((-1,-1)^T, \sigma_V^2 I) & f_3^V \sim N((1,-1)^T, \sigma_V^2 I) \\ f_1^Z \sim N((0,-5)^T, \sigma_Z^2 I) & f_2^Z \sim N((2,5)^T, \sigma_Z^2 I) & f_3^Z \sim N((-2,5)^T, \sigma_Z^2 I) \end{array}$$

with $(\alpha_1^V, \alpha_2^V, \alpha_3^V) = (0.5, 0.25, 0.25)$ and $(\alpha_1^Z, \alpha_2^Z, \alpha_3^Z) = (0.5, 0.25, 0.25)$. The categorical explanatory variable X is defined as the mixture component from which V is generated. Notice that this induces three groups whose within dispersion is related to the standard deviation σ_V. Thus, Y is related to X if the association parameter π assumes values close to 1. If π assumes low values, Y does not depend upon X (through V) but, rather, upon Z. In particular, the value $\pi = 0$ in this model corresponds to the null hypothesis of *no association*. This null hypothesis consists of the fact that the groups induced by V have no explanatory power on Y.
We conducted some simulations to explore the relationship between π and the IPG. For fixed values of σ_V^2, σ_Z^2 we repeatedly generated samples of size N from the model above, and estimated the expected value of the IGP measure over the simulated samples. We used 1000 simulated datasets for each value of π. As an illustration, the left panel in Figure 1 shows the monotonicity that was observed across the simulations (results refer to the case $\sigma_Z^2 = 5$; similar patterns were observed for different values). This behavior suggests that *IGP* may be considered a reasonable measure of dependence.
However, a confounding effect exists in general between association (as measured by π) and the strength of the structure in Y. For example, if $\pi = 1$ but Y has weak structure (equivalently, if Y coincides with V but the variance σ_V^2 is very large) then the groups induced by V will not retain information on the dispersion of Y. This situation will practically coincide with the case of no association, even though $\pi = 1$. This behavior appears to be a general characteristic of this problem in general, and should be kept in mind when interpreting the index.
In other words, the ability of the IGP to measure the level of association is dependent on the fact that there is some structure in Y to begin with. If Y has no structure, so that the X-groups can essentially be viewed as a random selection from the observations' labels then *any* measure of association will be useless. By construction IGP assumes values ranging from 0 to 1. As mentioned above, the maximum value is reached only if: (i) There is a strong association between Y and X; and (ii) Y has a strong structure, i.e., Y

Figure 1: *$E(IGP)$ (left) and power (right) estimated over 1000 simulations. Plots are based on samples of size $N = 100$ (please refer to the text for details) .*



can be meaningfully partitioned into clusters having a low within-dispersion. If attention is focused on the evaluation of association, we may be interested in removing the dependence of the index (or, better, of its maximum value) on the Y-structure. Consistently with what we pointed out above, in simulations we observed that when the Y structure is highly dispersed, the IGP index does not reach its theoretical maximum value (one) and, moreover, the index shows a low sensitivity to the strength of association (i.e., it increases very slowly as π increases). This poses the problem of finding sharp bounds for the IGP. Therefore, the capability of X to describe the Y-dispersion should not be evaluated in absolute terms, i.e. by comparing the X-groups with an hypothetical optimal partition having IGP equal to one, as such a partition might simply not exist. On the contrary, P_X should be compared with its best competitor, say $P^* = P_{X^*}$. If $IGP(Y|X^*) << 1$, then $IGP(Y|X)$ should be compared not with one, but with $IGP(Y|X^*)$. This raises the problem of determining X^* or, better, the maximum attainable IGP value. One possibility is to limit attention to the class of partitions having the same structure as P_X (i.e., having the number of groups and group sizes as P_X). Adaptive optimization techniques (e.g., genetic algorithms) or search algorithms (e.g., the greedy heuristics proposed in Kalantari *et al.* (1993)) may perform adequately, as enumeration algorithms are clearly computationally prohibitive.

3 Inferential aspects

We now discuss some distributional and inferential aspects of the IGP measure as used here. Firstly, note that the IGP is a sample average and that it is therefore a (strongly) consistent estimator of $Pr(\{Y \text{ and } NN(Y) \text{ belong to the same group induced by } X\})$ as N tends to infinity.

Under the null hypothesis of no association ($\pi = 0$ in our model) it can be shown that $E[IGP(Y|X)] = \sum_{k=1}^{K}[Pr(X = x_k^*)]^2$. This value can be computed exactly for the simulated model above from the theoretical parameters α_k^V. For example, for the parameter values that were used one finds that $E[IGP(Y|X)] = .375$ under H_0. (This null value can be noted

in the left panel of Figure 1). On actual data, the quantity $E[IGP(Y|X)]$ under H_0 can be estimated from the observed counts in the K groups induced by X.
To test H_0 one can use a permutation distribution approach, i.e. extract random permutations from the set of the N X-group labels associated to the Y-observations. For each permutation of the labels the IGP is computed, and the p-value for $IGP(Y|X)$ is obtained as the proportion of IGP values that are more extreme (larger) than the observed $IGP(Y|X)$. A small p-value indicates rejection of H_0 in favor of the alternative hypothesis of association. To evaluate the power of this procedure one can simulate many datasets, and for each determine whether the permutation test would reject H_0 at a chosen alpha level. Thus one can easily estimate the power of the test to reject H_0 for different values of σ_V^2 and σ_Z^2, for various alternative values of π. Note that the rejection probability that one obtains with this procedure is averaged over all the possible group label counts that could be observed when distributing N observations over K groups. In other words, in our model the average is taken over a multinomial distribution having parameters $(N,(\alpha_V^1,\alpha_V^2,\alpha_V^3))$. In Figure 1 (right panel) the estimated powers of permutation tests are reported for the case when $\alpha = 0.1$ for various combinations of values of π and σ_V^2 (results refer to the case $\sigma_Z^2 = 5$; similar patterns were observed for different values). It is worth noting that the power appears to be increasing with π but its maximum value depends upon the dispersion within Y. This phenomenon is consistent with the discussion above on the confounding effect of π and the variance of Y.

4 Conclusions

In this paper we discuss the use of the IGP as a measure of association. This seems a promising direction. As any other measure of X/Y–association, the IGP reflects both dependency and the amount of "explainable" structure in Y. Hence, the rejection of the null hypothesis strongly suggests the existence of association.
This approach is very flexible, as it only requires the distances (or dissimilarities) between all possible pairs of cases, the dissimilarity being defined on the basis of Y only. Also, the procedure can be applied whatever the measure used to obtain the dissimilarities: for example, it is possible to consider time series (one for each case), sequence data (e.g. categorical time series or genetic sequences), and other situations where Y is complex but a dissimilarity measure between two cases can be defined.
Lastly, as we have pointed out, the null hypothesis considered above is a translation of the null hypothesis of ANOVA into this new context. Should the null hypothesis be rejected, it could be of interest to investigate further, evaluating which X-groups are responsible for the rejection using some adaptation of the post-hoc tests approach.

References

Kalantari B., Lari I., Rizzi A. and Simeone B. (1993) Sharp bounds for the maximum of the chi-square index in a class of contingency tables with given marginals, *Computational Statistics and Data Analysis*, 16, 19–34.

Kapp A. and Tibshirani R. (2007) Are clusters found in one dataset present in another dataset?, *Biostatistics*, 8, 9–31.